# Embedding strategies for effective use of information from multiple sequence alignments

Steven Henikoff[1] and Jorja G. Henikoff

[1]Howard Hughes Medical Institute

Basic Sciences Division

Fred Hutchinson Cancer Research Center

Seattle, Washington 98104

Corresponding Author: Steven Henikoff

Phone: (206) 667-4515

FAX: (206) 667-5889

E-mail: henikoff@howard.fhcrc.org

Running title: Embedding strategies for database searching

Manuscript contains 20 pages, 3 tables plus 6 figures

**Abstract**
We describe a new strategy for utilizing multiple sequence alignment information to detect distant relationships in searches of sequence databases. A single sequence representing a protein family is enriched by replacing conserved regions with position-specific scoring matrices (PSSMs) or consensus residues derived from multiple alignments of family members. In comprehensive tests of these and other family representations, PSSM-embedded queries produced the best results overall when used with a special version of the Smith-Waterman searching algorithm. Moreover, embedding consensus residues instead of PSSMs improved performance with readily available single sequence query searching programs, such as BLAST and FASTA. Embedding PSSMs or consensus residues into a representative sequence improves searching performance by extracting multiple alignment information from motif regions while retaining single sequence information where alignment is uncertain.

Keywords: multiple sequence alignment; homology searching; sequence databanks; consensus sequence; protein blocks

Improvements in the efficiency of large-scale DNA sequencing are leading to rapid increases in the number of databank sequences that lack genetic or biochemical documentation. This is clearly the case for databases of cDNA sequence fragments (Boguski *et al.*, 1993), which are thought to represent the majority of all human protein sequences, and for databases from large genome sequencing projects, such as the sequencing of uncharacterized bacterial genomes (Nowak, 1995). Matching these unknown sequences with sequences of known function is a major goal of genome research. Meanwhile, there remains the traditional goal of detecting homologs to help understand the function of a protein of interest to a biologist. Improved methods for detecting homology in database searches aid in achieving both goals.

It is widely assumed that homology detection can be improved by utilizing multiple alignment information. Either a single sequence query is used to search for homologs in a database of multiple sequence alignments (Henikoff,S & Henikoff, 1991, Attwood & Beck, 1994, Sonnhammer & Kahn, 1994) or patterns (Smith,RF & Smith, 1990, Bairoch, 1992), or an alignment or pattern query is used to search a sequence database (Gribskov *et al.*, 1987, Henikoff,S *et al.*, 1990, Neuwald & Green, 1994, Tatusov *et al.*, 1994, Krogh *et al.*, 1994, Thompson *et al.*, 1994b). Local motif-based methods typically employ ungapped alignments corresponding to the most conserved regions (see Henikoff,S, 1995 for a review), whereas profile and hidden Markov model (HMM) methods employ more global gapped alignments (Gribskov *et al.*, 1990, Krogh *et al.*, 1994, Eddy, 1996). In either case, position-specific scoring matrices (PSSMs) can represent all available information in a multiple sequence alignment, and several improvements in constructing PSSMs have recently been introduced (Brown *et al.*, 1993, Tatusov *et al.*, 1994, Henikoff,JG & Henikoff, 1996, Bailey & Gribskov, 1996, Sjolander *et al.*, 1996). However, there are no comprehensive evaluation studies that demonstrate the superiority of any multiple alignment-based querying method over single sequence querying methods such as BLAST (Altschul *et al.*, 1990), FASTA (Pearson, 1990) and Smith-Waterman (Smith,TF & Waterman, 1981). In the absence of such evidence, it remains possible that using all of the alignment information available at a position will degrade performance, especially for very diverse groups of proteins. Of particular concern are regions where multiple alignment is uncertain, such as outside of conserved regions.

Here we describe a new strategy for using multiple alignment information that combines the advantages of both motif-based and global methods, and we demonstrate that using multiple alignment information in this way greatly improves search results. A representative member of the protein group is chosen, and either PSSMs or consensus residues corresponding to conserved regions are embedded into it. In this way, multiple alignment information provides specificity for conserved regions while single sequence information is retained for regions of uncertain alignment. Comprehensive evaluations of various queries in database searching tests revealed that embedded queries performed much better than the best single unembedded sequence methods. The best PSSM methods (Tatusov *et al.*, 1994, Henikoff,JG & Henikoff, 1996) performed best, although a simple consensus embedding procedure outperformed the most widely used PSSM construction method. We conclude that multiple alignment information can be efficiently utilized to improve searching performance either directly, as in PSSM-embedding, or indirectly, as in consensus-embedding.

## Results

*Rationale*

Current methods for using multiple alignment information to search databases fall into two general categories: global and motif-based. Global methods use full multiple alignments including gaps, whereas motif-based methods use only conserved regions, typically without gaps, called blocks (Posfai *et al.*, 1989). There are potential drawbacks to both strategies, as illustrated by an example (Fig. 1). We selected two sequences from the helix-loop-helix (HLH) family of regulatory proteins, MYOD_CHICK and LYL1_HUMAN, plus sets of four other HLH sequences at random, and performed both global and motif-based alignments on all six sequences. This procedure was carried out 40 times with different random selections to assess variability in the results. Global alignment of the HLH domain of these proteins resulted in consistent alignment between MYOD_CHICK and LYL1_HUMAN in the helix 1 and helix 2 regions of the HLH domain, but variable alignment in the loop region, spanning 8-9 residues in the two proteins (Fig. 1A). A database searching query derived from an alignment is based on residue frequencies for each alignment position. Because alignment is highly uncertain in the loop region, the resulting frequencies there are of questionable value. For example, sometimes the asparagine residue in the loop region of MYOD_CHICK is aligned with an aspartic acid in LYL1_HUMAN and sometimes with a proline. It is typical that diverse families share only a few conserved regions, separated by extensive regions that are essentially unalignable. The rationale for basing queries on global multiple alignments is to increase sensitivity (detection of true positive sequences) by making use of more sequence information. However, treating residue frequencies within regions where alignment is uncertain or undefined the same as those within conserved regions is likely to decrease selectivity (avoidance of false positive sequences) rather than increase sensitivity.

Motif-based strategies attempt to increase selectivity by excluding regions of uncertain alignment and using only the conserved regions or blocks. However, the boundaries of the blocks can be difficult to establish and depend on the particular family representatives used to make them, potentially reducing sensitivity. For the HLH example, we used motif-based methods to find sets of blocks including MYOD_CHICK, LYL1_HUMAN and four other HLH sequences. Although alignments were identical in every case, block boundaries varied considerably, as shown for MYOD_CHICK in Fig. 1B. Perhaps more importantly, in about 30% of the trials, only one of the two blocks was found. This example illustrates a potential drawback to motif-based methods: Depending on the set of sequences presented to them, they sometimes miss conserved regions.

In the general strategy introduced here, blocks are embedded into a single sequence to represent a protein family for database searching. That is, regions between the blocks, where alignment is uncertain, are taken from a single sequence. This avoids using misaligned positions that compromise global strategies by reducing selectivity. Furthermore, because the single sequence includes any potentially informative region, it preserves sensitivity that might have been lost in the block-making process. Below, we describe and evaluate specific implementations of this embedding strategy.

*Evaluation procedures*

Embedding requires a set of conserved regions for a protein family and a single sequence representative. We obtained conserved regions for 249 different protein families from the Blocks Database of multiple alignments (Henikoff,S & Henikoff, 1993). For each family we determined a consensus derived from the conserved regions (see Methods) and selected a sequence closest to the consensus. Searches were done using BLAST (Altschul *et al.*, 1990), FASTA (Pearson, 1990) and SWAT (P. Green, personal communication), an implementation of the Smith-Waterman algorithm (Smith,TF & Waterman, 1981), with substitution matrices and gap penalties optimized for best performance (Pearson, 1995). Lists of true positive sequences for each protein family were taken from the PROSITE (Bairoch, 1992) entries corresponding to the blocks, but only sequences not present in the blocks were used for evaluation.

To validate our evaluation procedure, we compared our searching results with those of Pearson (Pearson, 1995), who utilized PIR superfamilies rather than PROSITE groups to provide queries and lists of true positives. In spite of these and other procedural differences, our results were very similar to his (Fig. 2): Using optimized parameters, FASTA performed slightly worse than SWAT (z=0.9) and BLAST (z=0.3), and with log-length correction of FASTA and SWAT scores, FASTA was significantly worse than SWAT (z=3.5), and significantly better than BLAST (z=2.6). The similarity of our results to those of Pearson confirms that our evaluation procedure works well for directly comparing pairs of competing sequences in database searching.

*Searching performance using single sequences*

We next compared the performance of the closest member sequence relative to the performance of the farthest member. With all three searching algorithms, much better performance was obtained with the closest than with the farthest member sequence (z=5.3 to 7.1, Fig. 3). Consistent performance differences were seen regardless of the method used to measure search results.

The best single sequence searching method combined the closest member query and SWAT with log-length score correction (Pearson, 1995). We compared database search results using this method with those using different types of queries constructed from multiple alignment information.

*Embedded position-specific scoring matrices perform well*

Multiple sequence alignments can be represented as position-specific scoring matrices (PSSMs), consisting of columns of scores for each amino acid derived from corresponding alignment columns. Previously, we tested several different methods for constructing PSSMs corresponding to blocks of conserved regions in groups of proteins (Henikoff,JG & Henikoff, 1996), confirming and extending the results of others (Tatusov *et al.*, 1994). Log-odds PSSMs incorporating position-based sequence weights and substitution-probability pseudo-counts performed best in those searching tests. To compare the searching performance of PSSM queries against single sequence queries, we used two programs that query a sequence database with a PSSM: BLIMPS/MULTIMAT and SWAT. BLIMPS queries a database with multiple independent PSSMs representing blocks of conserved regions for a group, then MULTIMAT combines the search results, scoring hits based on the order and distance between blocks in a

database sequence as well as on individual block scores (Henikoff,S *et al.*, 1995). MULTIMAT performance using log-odds PSSMs was significantly better than the best single sequence method (z=4.0, Fig. 4). MULTIMAT failures could be attributed to true positive fragments that did not contain any of the blocks, or to uninformative blocks for the group, a finding also reported by Wu *et al.* (Wu *et al.*, 1996).

SWAT applies the Smith-Waterman algorithm and optionally scores with a PSSM in place of a substitution matrix, with fixed gap penalties. We took the same PSSMs made from the blocks representing a group searched independently by BLIMPS and embedded them in the closest member sequence. Between and flanking the blocks, we used identically-scaled BLOSUM scores from the matrix used to generate PSSM pseudo-counts. As a consequence, residues between the blocks were scored as for the usual Smith-Waterman search, but those within the conserved regions were scored more specifically for the group. By using full-length sequences, SWAT should be much less susceptible than MULTIMAT to failures attributable to fragments or deficient blocks. We also embedded PSSMs made using the average score method (Gribskov *et al.*, 1987).

The effect of embedding on the distribution of search scores is illustrated for the HLH domain. The two log-odds PSSMs for Blocks BL00038A and BL00038B, which represent the basic-helix-loop-helix domain, were embedded into the closest member (MYOD_CHICK). In the embedded regions, the most conserved position (137) scores +12 for lysine, and from -8 to -15 for all other residues (Fig. 5A). Lys137 is invariant in the PROSITE pattern for this group (PS00038). A much less conserved position (143, an X residue in the PS00038 pattern but a lysine in the consensus), scores +6 for lysine, and from +4 to -7 for all other residues, still somewhat more specific than for unembedded sequence, where BLOSUM 55 scores a lysine match +6 and mismatches from +3 to -4. Graphical display of the overall score distribution for the embedded MYOD_CHICK sequence reveals a shift to strongly negative scores for non-conserved residues, increasing specificity within the embedded blocks relative to flanking and interblock regions scored by BLOSUM 55 (Fig. 5B).

SWAT performance using embedded PSSMs outperformed the best single sequence method very strongly with log-odds PSSMs (z=9.8, Fig. 4) but only slightly with average score PSSMs (z=0.7, Fig. 4). SWAT out-performed MULTIMAT using the same log-odds PSSMs (z=2.9), and this was the best query and searching algorithm combination in our tests.

*Consensus embedding improves single sequence performance*

Although queries with embedded log-odds PSSMs performed the best, most biologists do not have easy access to the special searching programs they require. However, a single consensus-embedded sequence can be constructed for use as query with standard searching programs. To make a suitable consensus, the same multiple alignment blocks used to construct PSSMs for embedding were used to choose consensus residues for similar embedding into the closest member. Each consensus-embedded query was tested for BLAST, FASTA and SWAT searching performance. For all three programs, consensus-embedded closest members strongly out-performed the corresponding set of unembedded sequences (z=7.4, 5.9, 6.9, Fig. 6). Leaving real sequence residues between the embedded regions is crucial, because excising these residues by replacing them with Xs (Patthy, 1987, Worley *et al.*, 1995) resulted in much worse performance (z=5.6, Fig. 6).

Consensus-embedded queries performed much better than PSSM-embedded queries with SWAT when PSSMs were constructed using the average score method (z=6.4, Fig. 6), although they performed much worse than PSSM-embedded queries when PSSMs were constructed using the log-odds method (z=8.3, Fig. 6).

Two different substitution matrices were tested for choosing consensus residues, with no significant performance differences for either BLAST or FASTA (data not shown). We also tested the highest log-odds PSSM score in each block position as the consensus residue, but queries embedded with these residues performed slightly less well than those using substitution matrices (data not shown); perhaps this is because single sequence searching programs use a substitution matrix instead of a PSSM, making it more advantageous to bias the query with these matrices. The closest member embedded with consensus residues that have been selected using a good substitution matrix should be adequate for use with diverse searching programs.

To judge these findings from the perspective of other improvements in sequence searching tools, we compared our results with performance differences found in evaluations of substitution matrices (Henikoff,S & Henikoff, 1993). The better performance of BLAST with its current default matrix, BLOSUM 62, over one of the worst of the tested matrices, PAM 250 (z=7.1, Fig. 6), is about the same as that seen for consensus embedded queries compared with unembedded queries using BLAST (z=7.4, Fig. 6). Only when we tested the +6/-1 matrix, which scores all matches +6 and all mismatches -1, was the performance improvement using BLOSUM 62 (z=10.8, Fig. 6) greater than that seen for PSSM-embedded queries compared with unembedded queries. We conclude that performance improvements using queries embedded with multiple alignment information compare very favorably with other major advances in homology searching.

*Simple patterns perform poorly*

Each PROSITE group in our test set includes one or more multiple alignment-based pattern determined following a semi-manual procedure (Bairoch, 1992). Using one PROSITE pattern representing each group to search SWISS-PROT and scoring a hit as any match to the pattern, we found that performance was much worse than that using unembedded closest member queries with SWAT (z=6.0, Fig. 4). Even the farthest member queries performed slightly better than patterns in our tests (z=0.8, Fig. 4). Therefore, a standard database search using a single sequence representative of a group can be expected to detect new members better than a pattern.

*Searching current databanks with consensus-embedded queries*

An example demonstrates the practical value of using consensus-embedded queries in searching sequence databanks with the BLAST Server. Inteins are protein introns that are spliced out post-translationally and are often found in unexpected locations (Cooper & Stevens, 1995). Because they belong to a diverged protein group, inteins can sometimes be detected by careful database searching methods (Pietrokovski, 1994). We used the set of 6 conserved blocks representing 8 inteins from the PRINTS database (Attwood & Beck, 1994) to embed into the closest member, Psp GB-D pol, an (unspliced) DNA polymerase containing an intein. Searches of the non-redundant protein database on the NCBI BLAST server using

default parameters revealed the presence of 8 inteins detected by the consensus-embedded sequence that were not detected by the unembedded sequence (Table 1). In addition, two confirmed inteins were detected at a higher level of significance. No inteins detected by the unembedded sequence were missed using the consensus-embedded sequence, nor were any detected by the consensus-embedded sequence at a lower level of significance. Better sensitivity was not accompanied by an increase in background: in fact BLAST detected 29 false positives at P<0.99 using the unembedded sequence query but only 16 using the consensus-embedded sequence. Just 2 of the 16 presumed false positives increased in significance with the consensus-embedded query, suggesting that improvements due to embedding provide useful criteria for distinguishing true positives from false positives within the twilight zone. The same rationale underlies the BLAST3 searching program, in which the involvement of a conserved region in a high-scoring multiple alignment can improve the significance of a twilight zone hit (Altschul & Lipman, 1990). In all, 13 of the 14 databank sequences considered to be inteins were reported using the consensus-embedded minimal sequence, compared to only 5 using the unembedded sequence as query. Especially interesting is the detection of Ssp dnaB (P=0.19), which had been overlooked until its recent discovery by S. Pietrokovski (Pietrokovski, 1996) using BLIMPS with MULTIMAT. Note that the use of PRINTS multiple alignments, which were constructed in an entirely different manner from the Blocks Database blocks used for evaluations, indicates that our strategy is a general one. We expect that consensus embedding will provide comparably improved performance employing protein or DNA multiple alignments generated by any means.

**Discussion**

Previous methods for utilizing multiple alignment information in database searching have been either global or motif-based. Global profile queries (Gribskov *et al.*, 1987, Eddy, 1996) attempt to increase sensitivity by including weak information from non-conserved regions, but these regions are difficult to align with confidence. Consequently, scores for conserved regions are diluted by scores for regions where alignment is uncertain. Perhaps because of this problem, global profile and hidden Markov model methods have been reported to perform less well than some motif-based methods (Wu *et al.*, 1996). This problem might also explain the observation that excision of regions of uncertain alignment can improve Smith-Waterman profile searching (Thompson *et al.*, 1994b). Motif-based queries attempt to increase selectivity by discarding weak information outside conserved regions in a multiple alignment. However, some of the discarded information might prove useful, and the definition of a motif can impose an undesirable length limitation on an alignment. Indeed our tests show that simple pattern representations of motifs do less well overall than searching with individual sequences using modern searching programs.

Embedding is an alternative strategy that overcomes the problem of uncertain alignment in global strategies as well as that of missing regions in motif-based strategies. Embedding is conceptually simple, and it leads to overall performance improvements comparable to other major advances in sequence homology detection. Dramatically improved searching performance results when either PSSMs or consensus residues are embedded into a representative sequence. The use of the same substitution matrix model for scoring both embedded and unembedded regions naturally balances the relative contributions of conserved

and diverged regions to the overall score.

Consensus embedding can be of great practical value to the biologist because it provides much better performance using existing searching systems and is easily understood. Although searching strategies that use embedded log-odds PSSMs (but not average score PSSMs) perform even better, they require special searching programs and can be computationally impractical for large-scale use. Consensus embedding may be the only multiple alignment-based method that a biologist can use to improve the searching performance of a query sent to proprietary databanks (Boguski *et al.*, 1993), which offer only single sequence searching methods. Currently, consensus embedding is available from the Blocks World-Wide Web server (http://blocks.fhcrc.org), which will automatically search the non-redundant protein databank via the BLAST server (http://ncbi.nlm.nih.gov) using queries derived from either PROSITE or user-defined protein families.

Full-length consensus sequences made from global multiple alignments have been used for homology searching (Sonnhammer & Kahn, 1994, Henikoff,S *et al.*, 1988, Patthy, 1987). However, the strong performance of consensus embedding suggests that biasing only the conserved regions of a representative sequence while leaving the rest of it unmodified is more effective. This finding supports the view of families of protein sequences as blocks of conserved regions separated by variable regions (Posfai *et al.*, 1989).

The value of embedding strategies should increase as databanks expand with sequences from large-scale projects, because there will be more alignable sequences representing any given group. The strategy is readily automated, and might be combined with iterative searching methods (Yi & Lander, 1994; Tatusov *et al.*, 1994) to provide an even more powerful automated system. Embedding strategies are general, and therefore can be extended to other situations in which motif-based alignment information is available for a group of sequences. In addition to searching sequence databases as described here, PSSM- and consensus-embedding strategies should be applicable to searching multiple alignment databases, as well as to methods used to obtain multiple sequence alignments.

## Methods

*Database searches*

Queries for searching were taken from protein groups represented in BLOCKS 5.0 (Henikoff,S & Henikoff, 1991) which is keyed to PROSITE 9.0 (Bairoch, 1992) and SWISS-PROT 22 (Bairoch & Boeckmann, 1992). The target database was SWISS-PROT 32, and the list of known true positive sequences in SWISS-PROT 32 was taken from the corresponding PROSITE 13.0.

The searching programs used were BLAST 1.4.7 (Altschul *et al.*, 1990) with default parameters, including the BLOSUM 62 substitution matrix (Henikoff,S & Henikoff, 1992); FASTA 1.7.a4 (Pearson, 1990) with ktup=1, optimization, the BLOSUM 50 matrix and gap penalties of (-12,-2); SWAT (P. Green, personal communication), an implementation of the Smith-Waterman local alignment algorithm (Smith,TF & Waterman, 1981) with the BLOSUM 55 matrix and gap penalties of (-12,-2); BLIMPS 3.0.0 with the MULTIMAT analysis program (Henikoff,S *et al.*, 1995); and the Unix version of PATMAT (Wallace & Henikoff, 1992). The parameters for FASTA and SWAT were chosen according to Pearson's

recommendations (Pearson, 1995), and tests using these programs were evaluated using both raw and log-length corrected scores.

*Choosing query sequences*

Multiple alignment information for choosing queries was obtained from the Blocks Database (Henikoff,S & Henikoff, 1993), which consists of ungapped partial multiple alignments (blocks) for groups of protein sequences. A set of blocks represents consecutive conserved regions of the group. To reduce the contribution of redundant sequences to the block, each member sequence was weighted using the position-based method (Henikoff,S & Henikoff, 1994), and these weights were used to count the occurrence of each amino acid in each position. We had previously identified a set of 257 challenging protein groups in BLOCKS 5.0 (Henikoff,S & Henikoff, 1993) and 249 of those were used here (the others did not meet our revised evaluation criteria which require that there be true positives not in the blocks). The Blocks Database provided sets of automatically generated blocks for each of these groups, but not all of the known group members were included in the blocks. To select representative queries for each group, we first constructed a consensus for each block. At each position in the block, a consensus residue was chosen from among those in the position that provided the highest non-negative pairwise alignment score from a substitution matrix averaged over all of the residues at that position in the sequence-weighted multiple alignment. Restricting the selected residue to be among those that appear in the position is consistent with our log-odds scoring methods for PSSMs in general and with the use of BLOSUM matrices in particular, although other methods may not require it [*e.g.* Gribskov & Veretnik, 1996]. If the highest average pairwise score was negative, then the position was represented by an X (wild-card residue). This similarity score method is illustrated with an example in Table 2. BLOSUM 62 was used to determine consensus residues reported here, but we also tested BLOSUM 50. Two sequences were chosen as queries from among those in the blocks for each group: the closest member was a sequence with the fewest differences from the consensus, and the farthest member was a sequence with the most differences from the consensus.

*Constructing multiple alignment queries*

Three types of multiple alignment-based queries were tested for each group: patterns, PSSM-embedded sequences, and consensus-embedded sequences. A pattern is a string of single and multiple residues, X (wild-card) residues and variable gaps that describes a conserved motif, and each group in PROSITE is provided with one or more patterns. A set of blocks in the Blocks Database is identified with one PROSITE entry for the group, and this pattern was chosen as a query. For example, the cytosine methyltransferase family is represented by two entries in PROSITE 9.0 and has 6 blocks (BL00094A-BL00094F) in BLOCKS 5.0; PS00094 was the pattern tested for this group. For searching, a pattern is slid along each database sequence and any match to the pattern is scored as a hit. Patterns were searched against SWISS-PROT using PATMAT.

Position-specific scoring matrices (PSSMs) were constructed from each group's blocks as described (Henikoff,JG & Henikoff, 1996). Each column in a PSSM corresponds to a position in the block and has 20 scores representing the presence of each amino acid in the

position. For searching, a PSSM is slid along a database sequence and at each alignment the score for each amino acid in the sequence is looked up in the column of the PSSM with which it is aligned, and the scores for all the columns are added. Two different types of PSSMs were tested: One was constructed by the average score method (Thompson *et al.*, 1994b, Gribskov *et al.*, 1987, Luthy *et al.*, 1994), and the other by a log-odds method using position-based substitution-probability pseudo-counts (Table 3), which was shown to perform the best in previous tests (Henikoff,JG & Henikoff, 1996). Individual block PSSMs were searched against SWISS-PROT by BLIMPS, and search results for multiple blocks from the same group were combined by MULTIMAT.

In addition, a PSSM the length of the closest member sequence for each group was constructed as follows: All block PSSMs for the group were computed in the same third bit scale as BLOSUM 55 and embedded at the locations of the blocks in the sequence. Outside of the blocks, the PSSM scores were taken from the BLOSUM 55 scoring matrix (Fig. 5A,B). These PSSM-embedded queries were searched against SWISS-PROT using SWAT.

Consensus-embedded queries were constructed by embedding the consensus residues, selected as described above, into the closest member sequences at the locations of the blocks in the sequence. Outside of the blocks, the original sequence residues were used or were replaced with X (Fig. 5C). The frequency-weighted average amino acid substitution score was used to score alignments with X.

*Evaluation of results*

Lists of true positives were obtained from groups in PROSITE v. 13.0 which is based on SWISS-PROT 32. Compared with PROSITE v. 9.0, PROSITE v. 13.0 includes updates that increase the number of true positives. Performance was assessed by comparing pairs of rank-ordered results, ignoring true positives present in the blocks. This procedure assured that no sequence contributing to the construction of a query was counted in evaluating results.

For each search, results lists were analyzed based on three different performance measures to evaluate the ability to detect true positives above background, as previously described (Henikoff,JG & Henikoff, 1996). Briefly, the >99.9% measure is the number of true positives above the $99.9^{th}$ percentile level of true negatives, the equivalence number is the rank at which the number of positive errors equals the number of negative errors (Pearson, 1995), and ROC is the area under the step curve representing the plot of true positives as a function of true negatives. These different measures emphasize different features of the results distribution, and so improved performance with all three is considered to be evidence for improvement overall. Following Pearson (Pearson, 1995), we report a z-value based on applying the sign-rank test to paired comparisons, where z=2 corresponds to the 95% significance level.

*Availability*

COBBLER (COnsensus Biasing By Locally Embedding Residues) is a program that computes PSSMS or consensus amino acids from a set of blocks and embeds them into a sequence belonging to the group. COBBLER was written in the C programming language and compiled for UNIX operating systems, and is available from the authors at henikoff@howard.fhcrc.org. Consensus-embedded sequences that represent protein groups in the Blocks Database can be

retrieved from the Blocks World-Wide Web (WWW) server (http://blocks.fhcrc.org). COBBLER also has been incorporated into the Blockmaker WWW page and e-mail server (blockmaker@blocks.fhcrc.org). Blockmaker first runs the PROTOMAT system (Henikoff,S & Henikoff, 1991) on related unaligned sequences to obtain a set of blocks, and then COBBLER computes and embeds consensus residues into one of the sequences. This embedded sequence can be used to query a sequence databank.

**Acknowledgements**
This study was stimulated by a discussion with Peter Harte, and furthered by discussions at the 1996 Aspen Center for Physics Workshop on Identifying Features in Biological Sequences. We thank Phil Green for providing the SWAT program and for modifying it to score with PSSMs, and Shmuel Pietrokovski for analyzing the intein results. This work was supported by a grant from NIH (GM29009).

**References**

Altschul SF, Lipman DJ. 1990. Protein database searches for multiple alignments. *Proc Natl Acad Sci USA 87*:5509-5513.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Attwood TK, Beck ME. 1994. PRINTS-a protein motif fingerprint database. *Protein Engineering 7*:841-848.

Bailey TL, Gribskov M. 1996. The megaprior heuristic for discovering protein sequence patterns. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology,* Menlo Park, California: AAAI Press. pp 15-24.

Bairoch A. 1992. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res 20*:2013-2018.

Bairoch A, Boeckmann B. 1992. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res 20*:2019-2022.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST--Database for "expressed sequence tags". *Nature Gen 4*:332-333.

Brown MP, Hughey R, Krogh A, Mian IS, Sjolander K, Haussler D. 1993. Using Dirichlet mixture priors to derive hidden markov models for protein families. In: Hunter, L Searls, D Shavlik, J, eds. *Proceedings First International Conference on Intelligent Systems for Molecular Biology* Washington: AAAI Press. pp. 47-55.

Cooper AA, Stevens TH. 1995. Protein splicing: self-splicing of genetically mobile elements at the proteins level. *Trends Biochem Sci 20*:351-356.

Eddy SR. 1996. Hidden Markov models. *Curr Opin Struct Biol 6*:361-365.

Gribskov M, Veretnik S. 1996. Identification of Sequence Patterns with Profile Analysis. *Meth Enzymol 266*:198-212.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355-4358.

Gribskov M, Luthy R, Eisenberg D. 1990. Profile analysis. *Meth Enzymol 183*:146-159.

Henikoff JG, Henikoff S. 1996. Using substitution probabilities to improve position-specific scoring matrices. *CABIOS 12*:135-143.

Henikoff S. 1995. Comparative methods for identifying functional domains in protein sequences. In: El-Gewely RM, ed. *Biotechnology Annual Review, vol 1.* Amsterdam: Elsevier. pp 129-147.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res 19*:6565-6572.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA 89*:10915-10919.

Henikoff S, Henikoff JG. 1993. Performance evaluation of amino acid substitution matrices. *Proteins: Struct Funct Genet 17*:49-61.

Henikoff S, Henikoff JG. 1994. Position-based sequence weights. *J Mol Biol 243*:574-578.

Henikoff S, Haughn GW, Calvo JM, Wallace JC. 1988. A large family of bacterial activator proteins. *Proc Natl Acad Sci USA 85*:6602-6606.

Henikoff S, Wallace JC, Brown JP. 1990. Finding protein similarities with nucleotide sequence databases. *Meth Enzymol 183*:111-132.

Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S. 1995. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene 163*:GC17-GC26.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. *J Mol Biol 235*:1501-1531.

Luthy R, Xenarios I, Bucher P. 1994. Improving the sensitivity of the sequence profile

method. *Prot Sci 3*:139-146.

Neuwald AF, Green P. 1994. Detecting patterns in protein sequences. *J Mol Biol 239*:698-712.

Nowak R. 1995. Bacterial genome sequence bagged. *Science 269*:468-470.

Patthy L. 1987. Detecting homology of distantly related proteins with consensus sequences. *J Mol Biol 198*:567-577.

Pearson WR. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth Enzymol 183*:63-98.

Pearson WR. 1995. Comparison of methods for searching protein sequence databases. *Prot Sci 4*:1145-1160.

Pietrokovski S. 1994. Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Prot Sci 3*:2340-2350.

Pietrokovski S. 1996. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res 24*:3836-3845.

Posfai J, Bhagwat AS, Posfai G, Roberts RJ. 1989. Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res 17*:2421-2435.

Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *CABIOS 12*:327-345.

Smith RF, Smith TF. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA 87*:118-122.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195-197.

Sonnhammer ELL, Kahn D. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Prot Sci 3*:482-492.

Tatusov RL, Altschul SF, Koonin EV. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA 91*:12091-12095.

Thompson JD, Higgins DG, Gibson TJ. 1994a. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap

penalties and weight matrix choice. *Nucleic Acids Res 22*:4673-4680.

Thompson JD, Higgins DG, Gibson TJ. 1994b. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS 10*:19-29.

Wallace JC, Henikoff S. 1992. PATMAT: a searching and extraction program for sequence, pattern, and block queries and databases. *CABIOS 8*:249-254.

Worley KC, Wiese BA, Smith RF. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res 5*:173-184.

Wu CH, Zhao S, Chen H-L, Lo C-J, McLarty J. 1996. Motif identification neural design for rapid and sensitive protein family search. *CABIOS 12*:109-118.

Yi TM, Lander ES. 1994. Recognition of related proteins by iterative template refinement (ITR). *Prot Sci 3*:1315-1328.

**Table 1.** *BLAST[1] performance with a consensus-embedded intein query[2]*

| | | BLAST P-value | |
|---|---|---|---|
| NR protein ID | Host protein | PSU00707_1 | +Consensus-embbed |
| PSU00707_1[3] | Psp GB-D pol | 0 | 0 |
| PYWKODPOL_1[3] | Psp KOD pol | 0 | 0 |
| DPOL_THELI[3] | Tli pol | 0 | 0 |
| VATA_YEAST[3] | Sce VATA | - | 0.00031 |
| MGDNAGRA_1 | Myo gyrA | 0.030 | 0.0021 |
| MFDNAGRA_1 | Mfl gyrA | 0.65 | 0.0089 |
| RECA_MYCTU[3] | Mtu recA | - | 0.012 |
| HO_YEAST | HO endo | - | 0.049 |
| MLDNAGRA_1 | Mle gyrA | - | 0.16 |
| SYCSLLE_71 | Ssp dnaB | - | 0.19 |
| VATA_CANTR[3] | Ctr VATA | - | 0.56 |
| MKDNAGRA_1 | Mka gyrA | - | 0.75 |
| U00013_9[3] | Mle pps1 | - | 0.95 |
| RECA_MYCLE[3] | Mle recA | - | - |
| # false positives with P<0.99 | | 29 | 16 |

[1]BLASTP search of the "non-redundant" (NR) protein database with default parameters and sum statistics using the NCBI e-mail server on 2/3/96. Identical and nearly identical sequence entries have been pruned. Dash indicates that the sequence was not reported.
[2]PSU00707_1, the member from among the 8 intein-containing sequences documented in PRINTS (Attwood & Beck, 1994) closest to the consensus computed as in Table 1 from the 6 blocks in PRINTS for this group. Because this is a DNA polymerase II homolog, all DNA polymerases in the results list were ignored.
[3]Present in the PRINTS blocks.

**Table 2.** *Choosing consensus residues*

*A. Simple but real alignment of nitrogenase segments*

| | Swiss-Prot ID | AA# | Alignment | Sequence weights |
|---|---|---|---|---|
| 1 | NIFK_THIFE | 482 | YQGAV | 0.867 |
| 2 | NIFK_ANASP | 476 | YQGGL | 0.867 |
| 3 | NIFD_CLOPA | 176 | YKGVS | 1.000 |
| 4 | NIFD_AZOVI | 452 | FDGFA | 1.267 |
| | consensus | | YQGFA | |

*B. Sample calculation of pairwise similarity scores (for column 5)*

| AA | Column AA | Score[1] | Weight | Weighted score | Average score[2] |
|---|---|---|---|---|---|
| V | V | 4 | 0.867 | 3.486 | |
| | L | 1 | 0.867 | 0.867 | |
| | S | -2 | 1.000 | -2.000 | |
| | A | 0 | 1.267 | 0.000 | 0.588 |
| L | V | 1 | 0.867 | 0.867 | |
| | L | 4 | 0.867 | 3.486 | |
| | S | -2 | 1.000 | -2.000 | |
| | A | -1 | 1.267 | -1.267 | 0.272 |
| S | V | -2 | 0.867 | -1.734 | |
| | L | -2 | 0.867 | -1.734 | |
| | S | 4 | 1.000 | 4.000 | |
| | A | 1 | 1.267 | 1.267 | 0.450 |
| A | V | 0 | 0.867 | 0.000 | |
| | L | -1 | 0.867 | -0.867 | |
| | S | 1 | 1.000 | 1.000 | |
| | A | 4 | 1.267 | 5.068 | 1.300 |

[1]From the BLOSUM 62 scoring matrix.
[2]If the highest average score is negative, then the consensus residue is X.

**Table 3.** *Constructing a log-odds PSSM*

*A. Sample calculation of counts for column 1 of the alignment in Table 1A*[1]

$n_Y = 0.867 + 0.867 + 1.000 = 2.734$
$n_F = 1.267$
$n_A = 0.000$

*B. Pseudo-counts*[2]

$b_Y = 10 * (n_Y / 4) * (q_{YY} / Q_Y) + (n_F / 4) * (q_{FY} / Q_F) = 2.462$
$b_F = 10 * (n_Y / 4) * (q_{FY} / Q_Y) + (n_F / 4) * (q_{FF} / Q_F) = 2.130$
$b_A = 10 * (n_Y / 4) * (q_{AY} / Q_Y) + (n_F / 4) * (q_{AF} / Q_F) = 0.386$

*C. Odds-ratios*[3]

$r_Y = ((n_Y + b_Y)/(4+10)) / f_Y = ((2.734 + 2.462)/14) / 0.032 = 11.598$
$r_F = ((1.267 + 2.130)/14) / 0.040 = 6.066$
$r_A = ((0.000 + 0.386)/14) / 0.077 = 0.358$

*D. Third bit log odds-ratios used in PSSM*

$w_Y = 3 * \ln(r_Y) / \ln 2 = 10.610$
$w_F = 7.804$
$w_A = -4.447$

[1] Calculations for alanine (A) are shown as an example of those for the 18 amino acids that do not occur in this column of the block. The counts for tyrosine and phenylalanine (Y and F), which do occur, are computed from the sequence weights.

[2] q and Q are probabilties and marginal probabilities underlying the Blosum 62 amino acid substitution matrix. A total of 10 pseudo-counts is used for this column and distributed among the 20 amino acids. Note the large number of pseudo-counts relative to counts in this example in consequence of the small number of sequences observed.

[3] The f values are background amino acid frequencies taken from Swiss-Prot.

**Fig. 1.** Comparison of global and block-based methods in a region of alignment uncertainty. Partial sequences consisting of the HLH domain (Pfam PF00010, http://www.sanger.ac.uk) from members of the helix-loop-helix (HLH) family of regulatory proteins (PROSITE v. 5.0 PS00038) were used for multiple alignment. MYOD_CHICK was chosen to be the reference sequence (see text and Fig. 2) and LYL1_HUMAN was selected at random to serve as a comparison sequence. Sets of four other sequences were selected at random and all six sequences aligned using ClustalW v. 1.6 (Thompson *et al.*, 1994a) and Blockmaker (Henikoff,S *et al.*, 1995) from the BCM multiple alignment launcher (http://dot.imgen.bcm.tmc.edu). This procedure was repeated with sets of 4 randomly selected HLH sequences for a total of 40 trials. In 15 cases, neither GIBBS nor MOTIF provided 2 blocks consisting of all 6 sequences, and these Blockmaker trials were not used. A. Summary of 40 alignments between MYOD_CHICK and LYL1_HUMAN reported by ClustalW for the helix-loop-helix region when the other 4 sequences were varied. Spaces delimit the variable region, and dashes indicate gaps. B. Summary of 25 block boundaries depicted for MYOD_CHICK. The number of residues in the interblock region is given in parentheses. In both cases, percent occurrence of each alignment or block boundaries are shown.

**Fig. 2**. Evaluation of single sequence searching programs using the closest member sequence. Pairs of query sequences (closest and farthest members) representing 249 groups of proteins were searched against SWISS-PROT 32 using three different searching programs. Following Pearson (Pearson, 1995), results are based on equivalence number (E-value), and z-values for the sign-rank test are displayed (grey bars). The solid bars represent the number of groups for which the program on the right performed better than that on the left, the open bars represent the number of test sequences for which it performed worse, and ties are not shown. Where indicated, raw FASTA or SWAT scores were used rather than log-length corrected scores.

**Fig. 3.** Effect of query choice on searching performance. Pairs of 249 searches were done as in Fig. 2. The solid bars represent the number of groups for which the closest member sequence performed better than the farthest member sequence, the open bars represent the number of groups for which it performed worse, and ties are not shown. Three performance measures were employed: the number of true positives scoring above the $99.9^{th}$ percentile level of true negatives, E-value, and the ROC area.

**Fig. 4.** Evaluation of patterns and PSSM-embedding. Pairs of 249 searches were done as in Fig. 2. The solid bars represent the number of groups for which the query and program combination on the right performed better than the combination on the left, the open bars represent the number of groups for which it performed worse, and ties are not shown. Only E-values are shown. The query and program combinations are: PROSITE pattern with PATMAT (pattern); closest member sequence with SWAT (SWAT); farthest sequence member query with SWAT (SWAT (farthest)); log-odds PSSMs with BLIMPS/MULTIMAT (MULTIMAT); average score PSSMs embedded into the closest member sequence with SWAT (SWAT avg. score PSSM); log-odds PSSMs embedded into the closest member sequence with SWAT (SWAT PSSM), which scored alignments with embedded regions using the PSSMs rather than BLOSUM 55.

**Fig. 5.** Embedding PSSMs and consensus residues into the closest member sequence (MYOD_CHICK) for the HLH family. A) PSSM scores for position 137, where lysine is invariant (K 137), for position 143, where lysine is preferred over other residues (K 143), and BLOSUM 55 scores for lysine [k (B55)]. B) Distribution of scores along the PSSM-embedded sequence for residues 81-180, showing the location of blocks BL00038A-B along the sequence (black boxes), the location of the PROSITE pattern PS00038 (gray box) beginning at position 137:
K-[LIVMAG]-x-[IT]-[IL]-x(2)-[STAV]-x(2)-[YHV]-[LIVMA]-x(2)-[LIVM] (Bairoch, 1992) and the location of positions 137 and 143 illustrated in a) (arrows at bottom). Each symbol is the score for one of the 20 amino acids when aligned with the indicated position along the X-axis. C) The consensus residues determined for BL00038A-B were embedded (underlined segments), changing the residues shown in bold.

**Fig. 6.** Evaluation of consensus embedding, with substitution matrix comparisons as a reference. Pairs of 249 searches were done as in Fig. 2, and comparisons showing E-values are as in Fig. 4. The queries are: closest member sequence (closest); consensus residues embedded in the closest member sequence (consensus); consensus residues embedded in the closest member sequence with Xs around them (cons + X); average score PSSMs embedded into the closest member sequence (avg. score PSSM); log-odds PSSMs embedded into the closest member sequence (PSSM). SWAT scored alignments with embedded regions using the PSSMs rather than BLOSUM 55. The bars labelled "PAM250", "+6/-1" and "BLOSUM62" compare the performance of BLAST using two of those scoring matrices and the "closest" query as a reference for the degree of searching improvement observed in Figs. 2-5.

A. Alignment variability

| | | | Occurrence |
|---|---|---|---|
| KVNEAFETLKRCTS | -TNPNQRLP | KVEILRNAIRYIESL | 2.5% |
| NVNGAFAELRKLLP | THPPDRKLS | KNEVLRLAMKYIGFL | |
| KVNEAFETLKRCTS | TN-PNQRLP | KVEILRNAIRYIESL | 12.5% |
| NVNGAFAELRKLLP | THPPDRKLS | KNEVLRLAMKYIGFL | |
| KVNEAFETLKRCTS | TNP-NQRLP | KVEILRNAIRYIESL | 60% |
| NVNGAFAELRKLLP | THPPDRKLS | KNEVLRLAMKYIGFL | |
| KVNEAFETLKRCTS | TNPN-QRLP | KVEILRNAIRYIESL | 17.5% |
| NVNGAFAELRKLLP | THPPDRKLS | KNEVLRLAMKYIGFL | |
| KVNEAFETLKRCTS | TNPNQ-RLP | KVEILRNAIRYIESL | 5% |
| NVNGAFAELRKLLP | THPPDRKLS | KNEVLRLAMKYIGFL | |
| KVNEAFETLKRCTS | TNPNQRLP- | KVEILRNAIRYIESL | 2.5% |
| NVNGAFAELRKLLP | THPPDRKLS | KNEVLRLAMKYIGFL | |

B. Block variability

| | | | |
|---|---|---|---|
| KVNEAF | (11) | NQRLPKVEILRNAIRYIESL | 4% |
| KVNEAFETLK | (6) | PNQRLPKVEILRNAIRYIESL | 16% |
| KVNEAFETLK | (7) | NQRLPKVEILRNAIRYIESL | 24% |
| KVNEAFETLK | (9) | RLPKVEILRNAIRYIESL | 4% |
| KVNEAFETLK | (12) | KVEILRNAIRYIESL | 4% |
| KVNEAFETLKR | (5) | PNQRLPKVEILRNAIRYIESL | 4% |
| KVNEAFETLKR | (6) | NQRLPKVEILRNAIRYIESL | 8% |
| KVNEAFETLKRCTS | (2) | PNQRLPKVEILRNAIRYIESL | 8% |
| KVNEAFETLKRCTS | (3) | NQRLPKVEILRNAIRYIESL | 12% |
| KVNEAFETLKRCTST | (2) | NQRLPKVEILRNAIRYIESL | 12% |
| KVNEAFETLKRCTST | (4) | RLPKVEILRNAIRYIESL | 4% |

FIGURE 1

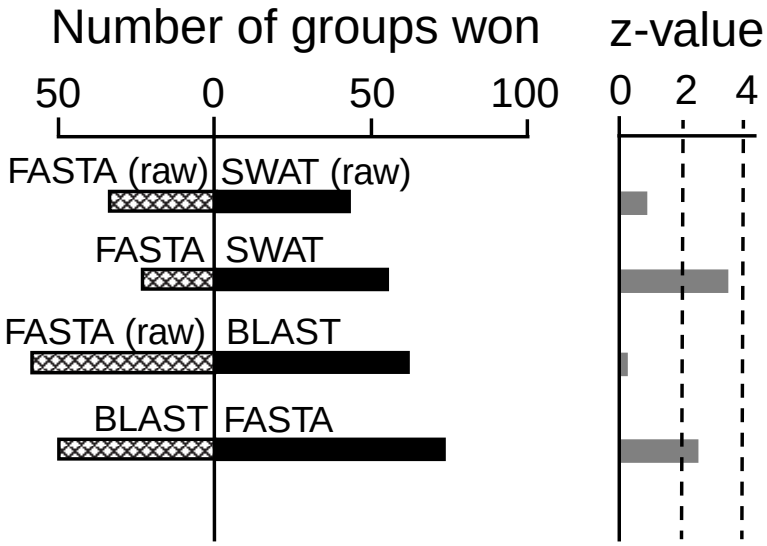

Number of groups won
50
0
50
100
FASTA (raw)
SWAT (raw)
FASTA
SWAT
FASTA (raw)
BLAST
BLAST
FASTA
z-value
0
2
4

FIGURE 2

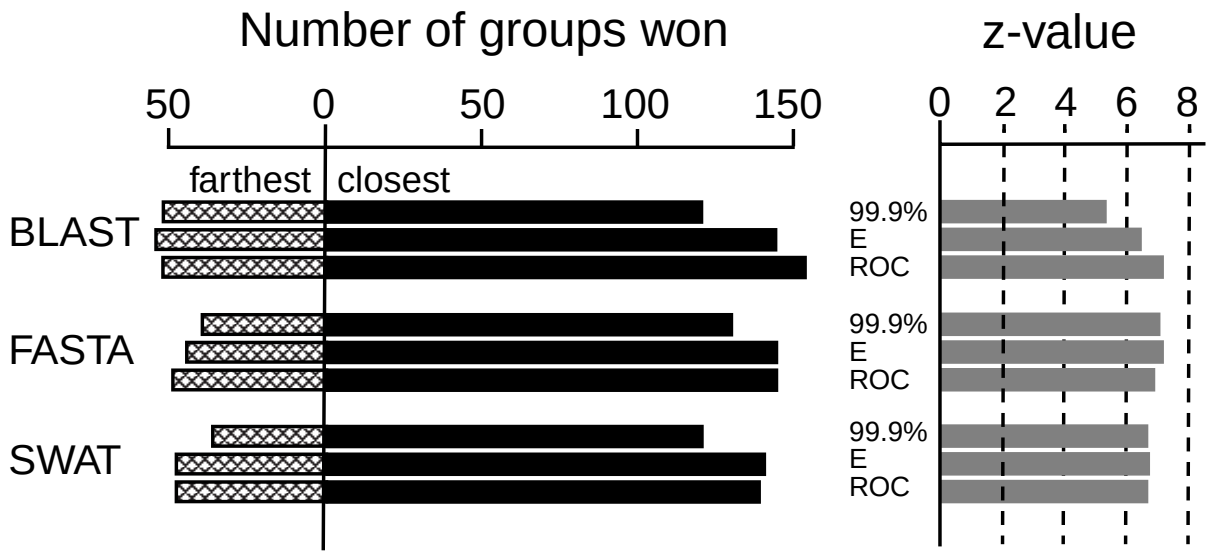

Number of groups won
50
0
50
100
150
farthest
closest
BLAST
FASTA
SWAT
z-value
0
2
4
6
8
99.9%
E
ROC
99.9%
E
ROC
99.9%
E
ROC

FIGURE 3

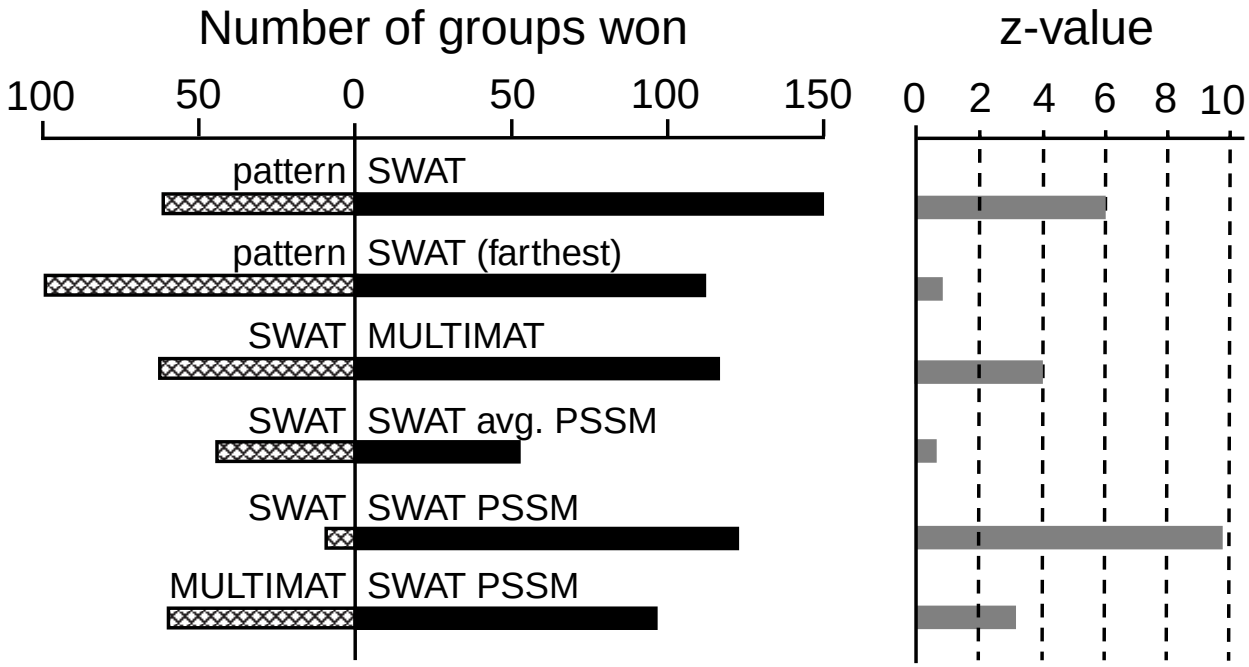
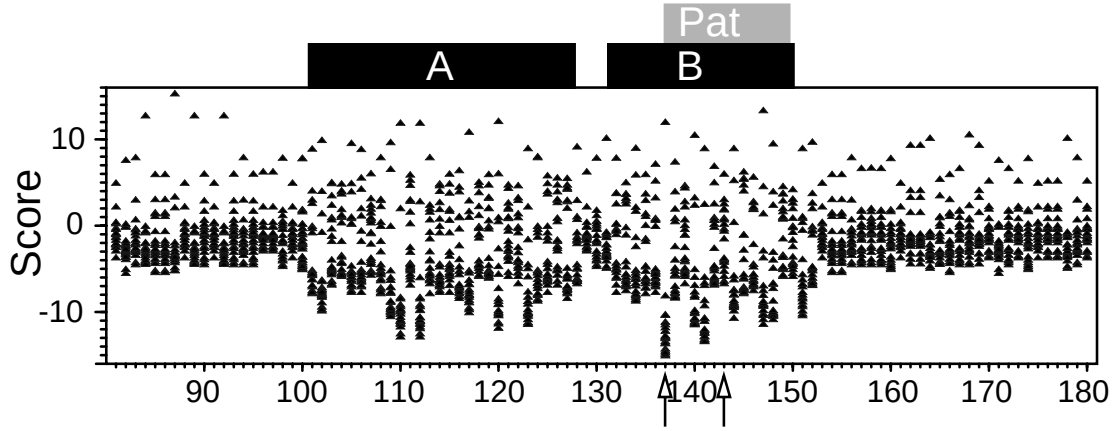

Number of groups won
100
50
0
50
100
150
pattern
SWAT
pattern
SWAT (farthest)
SWAT
MULTIMAT
SWAT
SWAT avg. PSSM
SWAT
SWAT PSSM
MULTIMAT
SWAT PSSM
z-value
0
2
4
6
8
10

FIGURE 4

A) Scores

| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K 137 | -13 | -14 | -12 | -10 | -15 | -13 | -11 | -14 | +12 | -15 | -13 | -11 | -13 | -10 | -8 | -12 | -13 | -14 | -15 | -14 |
| K 143 | +4 | -5 | -4 | 0 | -4 | -6 | +3 | -1 | +6 | +2 | +3 | +2 | -6 | +1 | +1 | 0 | -4 | -1 | -7 | -5 |
| k (B55) | -1 | -4 | -1 | +1 | -4 | -2 | 0 | -4 | +6 | -3 | -2 | 0 | -1 | +2 | +3 | 0 | -1 | -3 | -4 | -2 |

B) PSSM embedding (MYOD_CHICK)




C) Consensus embedding

mdllgpmemtegslcsftaaddfyddpcfntsdmhffedldprlvhvggl
lkaeehphtrapprepteeehvrapsghhqagrcllwackackrkttnad
RRK**QHN**MRERRR**RE**KVNEAF**XE**LK**KMIP**tnp**KK**LPKVEILR**K**A**VE**YI**Q**S
LQallreqedayypvlehysgesdassprsncsdgmmeysgppcssrrrn
sydssyytespndpkhgkssvvssldclssiveristdnstcpilppaea

FIGURE 5

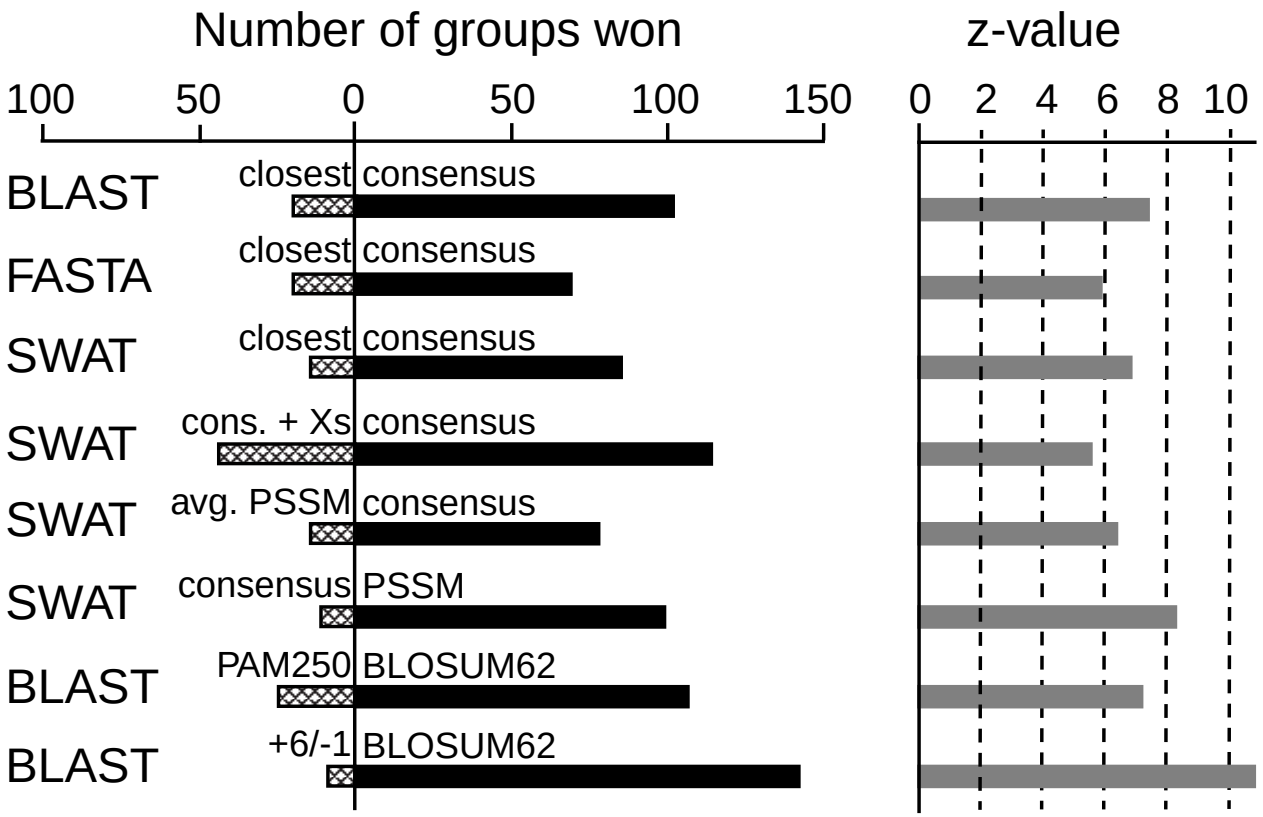

Number of groups won
z-value
100
50
0
50
100
150
0
2
4
6
8
10
BLAST
closest
consensus
FASTA
closest
consensus
SWAT
closest
consensus
SWAT
cons. + Xs
consensus
SWAT
avg. PSSM
consensus
SWAT
consensus
PSSM
BLAST
PAM250
BLOSUM62
BLAST
+6/-1
BLOSUM62

FIGURE 6